Concepts and methods for integrating language typology and sociolinguistics

Di Garbo, Francesca

Officinaventuno
2021

Di Garbo , F , Kashima , E , Napoleão de Souza , R & Sinnemäki , K 2021 , Concepts and methods for integrating language typology and sociolinguistics . in S Ballarè & G Inglese (eds) , Tipologia e Sociolinguistica : Verso un approccio integrato allo studio della variazione: Atti del Workshop della Società Linguistica Italiana 20 settembre 2020 . nuova serie , vol. 5 , Officinaventuno , Milano , pp. 143-176 , Societas Linguistica Europaea , 26/08/2020 . https://doi.org/10.17469/O2105SLI000005

http://hdl.handle.net/10138/334697
https://doi.org/10.17469/O2105SLI000005

DOI: 10.17469/O2105SLI000005

FRANCESCA DI GARBO, ERI KASHIMA,
RICARDO NAPOLEÃO DE SOUZA, KAIUS SINNEMÄKI[1]

Concepts and methods for integrating language typology and sociolinguistics

Questo articolo presenta le componenti costitutive di un programma di ricerca per lo studio tipologico dell'adattamento linguistico, ovvero del modo in cui le lingue cambiano in relazione ai contesti socio-storici e ambientali in cui sono utilizzate. Illustriamo una batteria di concetti e metodi volti a comparare sistematicamente contesti sociolinguistici e strutture linguistiche attraverso lo studio di comunità in contatto. Dimostriamo che questi concetti e metodi possono essere usati per studiare i correlati sociolinguistici della diversità e del mutamento linguistico in almeno tre modi: (1) per comprendere le cause del cambiamento linguistico, (2) per creare una base di dati rappresentativi di comunità, fattori sociolinguistici e linguistici, (3) per formulare generalizzazioni sulla base di studi comparativi a livello interculturale e interlinguistico.

This paper presents the building blocks of a comprehensive framework for the typological study of linguistic adaptation, i.e. how languages change in relation to the socio-historical and environmental contexts in which they are used. We showcase a battery of concepts and methods that are geared towards systematically comparing sociolinguistic environments and linguistic structures through the study of communities in social contact. We show that these concepts and methods can be used to investigate sociolinguistic correlates of linguistic diversity and language change in at least three ways: (1) to unravel causal factors related to language change, (2) to create datasets simultaneously addressing selection of communities, sociolinguistic features, and linguistic features, and (3) to formulate generalizations from empirically-grounded cross-cultural and cross-linguistic comparisons.

Parole chiave: adattamento linguistico, contatto linguistico, mutamento linguistico, tipologia, sociolinguistica comparativa

Keywords: linguistic adaptation, language contact, language change, typology, comparative sociolinguistics

[1] All authors equally contributed to this work. The ordering of authors is merely alphabetical.

1. *Introduction*

Research on the non-linguistic correlates of linguistic diversity has been on the rise over the past few decades. Studies from a variety of interconnected fields – such as language typology, sociolinguistics, and language evolution – have shown how patterns of language structures may change under the influence of the larger sociohistorical and environmental contexts in which languages are (or have been) used (see, among others, Wray & Grace 2007; Lupyan & Dale 2010; Trudgill 2011; Bentz & Winter 2013; Everett *et al.* 2015; Sinnemäki & Di Garbo 2018; Blasi *et al.* 2019). These processes of change are loosely captured under the umbrella term *linguistic adaptation* (Lupyan & Dale 2016).

Past studies on linguistic adaptation have individually tackled a diverse range of linguistic and non-linguistic features (e.g. phonemic inventory, morphological complexity for the former; population size, proportions of second language users, and climate for the latter), and investigated how these interact in processes of language change. The findings from these studies, as well as from related areas of research within linguistics, are indicating the need for a more holistic approach (cf. Hruschka *et al.* 2009). The time is thus ripe for establishing a common ground for the systematic study of linguistic adaptation while developing methodologies and tools that are specifically geared towards understand its impact on linguistic diversity and language evolution.

In this paper, we define linguistic adaptation as the processes whereby languages change in a way that enhances their learnability, efficiency of comprehension, and transmissibility in relation to the specific socio-historical and environmental contexts in which they are used. Communicative needs may differ across cultures, depending e.g. on social network structure and density or the amount of language contact. In the context of language contact, we can say that languages may adapt to being learned and used by a bi- or multilingual population. This definition of linguistic adaptation fits with the notion of adaptive changes in cultural evolution, which are defined as those that improve the transmissibility and frequency of a cultural trait (Lupyan & Dale 2016: 650).

Our aim is to contribute to advance research on linguistic adaptation by presenting the research framework that we developed

in the context of the ERC-funded project Linguistic Adaptation: Typological and Sociolinguistic Perspectives to Language Variation (GramAdapt, PI Kaius Sinnemäki). This framework is geared towards the task of comparing sociolinguistic environments and linguistic structures with one another through the in-depth study of communities in social contact. We thus approach linguistic adaptation in a narrow sense, that is by considering changes in language structures that are arguably the result of language contact and the influence of the wider sociolinguistic environment. While we do not assume that all instances of language change in contact situations are by default the result of adaptation processes, through the proposed research design we aim at providing a baseline against which hypotheses about linguistic adaptation can be tested in systematic and controlled ways. We developed the concepts and methods presented in this paper concurrently. The research design is articulated in five parts, each of which we explain below.

Firstly, we conducted a review of proposed explanatory factors for language change under contact situations since our goal is to better understand linguistic adaptation from the perspective of social contact between communities. In this literature review, we primarily consider patterns of socially motivated language variation and change that hinge on some degree of multilingual and bilingual language use because these are the defining features of language contact.

Secondly, we develop a typological approach for comparing sociolinguistic scenarios with one another. Because our goal is to study the relationship between languages and their sociolinguistic environments from a global perspective, we cannot always rely on naturalistic data. The main reason for this is poor availability of language corpora annotated for both linguistic and sociolinguistic features. Instead, our method is based on comparative tools devised by researchers, along the same lines as in state-of-the-art research in language typology.

Thirdly, we develop new methods for selecting sample languages and communities in contact from all around the world because we want to be able to compare linguistic structures and language contact scenarios on a global scale.

Fourthly, we develop a sociolinguistic questionnaire for collecting data on language contact scenarios. The questionnaire design is informed by established knowledge on the dynamics of language use

and language change in speech communities, by our sampling strategy, and by the conceptual tools we developed for comparing sociolinguistic contexts with one another. The sociolinguistic questionnaire is at the center of the data collection process implemented by the GramAdapt project.

The fifth and final part concerns the linguistic variables that we will use in order to test hypotheses about linguistic adaptation. These cover a range of domains of language structures spanning phonology, morphosyntax, and the lexicon. The languages of the sample will be coded for these variables, and their distribution will be then cross-checked with the sociolinguistic profiles emerging from the questionnaire data.

The different parts of the research design relate to each other, in such a way that choices in one subpart (e.g. explanatory factors) affect choices in the other subparts (e.g. questionnaire design). In this paper, we present each of the five parts of this research design starting with explanatory factors for contact-induced change (Section 2), comparative approaches to sociolinguistic environments (Section 3), and language sampling techniques for the investigation of contact scenarios (Section 4). We then illustrate the design principles and workings of the sociolinguistic questionnaire (Section 5) and an overview of the linguistic variables of choice (Section 6). Some concluding remarks are given in Section 7.

2. *Explanatory factors for contact-induced change*

One straightforward way to establish a common ground for studying linguistic adaptation is to explore the literature that addresses the topic. Given that language phenomena are intrinsically connected to other aspects of human activity, we set out to conduct a literature review on contact-induced change from a broad perspective. That is, instead of focusing on the contact literature alone, in this ongoing review we look at studies from several research domains, including psycholinguistics, bi-/multilingualism, second language learning, language acquisition, and areal linguistics.

In our review, we sought to uncover the more general patterns that may underlie contact-induced change. The choice of the specific areas mentioned above relates to existing mechanisms that have been pro-

posed in the literature to explain contact-induced change. We group these explanations here under the umbrella term *explanatory factors*.

Some of the individual explanatory factors that we identified in the first phase of our research include:

- Openness of the community (e.g. Dahl 2004; Wray & Grace 2007; Trudgill 2011)
- Geographic spread (e.g. Nichols 1992; Atkinson 2011)
- Population size (e.g. Hay & Bauer 2007)
- Number of linguistic neighbors (e.g. Lupyan & Dale 2010)
- Proportion of second-language (L2) speakers in a community (e.g. Bentz & Winter 2013)

Despite these many proposals, the overall picture of how linguistic structures change and diffuse in a community remains fragmentary. Most proposals usually explore a single aspect of the phenomenon without necessarily addressing the more fundamental questions behind the dynamics of contact-induced change. Many of the factors proposed in the literature are in fact proxies for phenomena that lead to structural change, and it remains unclear what their function when tested empirically is.

Some of the questions that the GramAdapt team addresses in the ongoing literature review are:

1. What are the cognitive mechanisms behind contact-induced change?
2. Which mechanisms apply to which changes?
3. What is the influence of social structure on linguistic structure?
4. Which types of social structure contribute to which changes?

More generally, this literature review focuses on gathering the various elements that have been proposed to explain contact-induced change. In order to achieve this, we established a principled way to conduct the literature review. For every source reviewed, we divided our task into two components: the *information extraction component*, and the *evaluation component*. The information extraction component was simply the search for specific pieces of information that relate to our topic of research. The evaluation component involved our own assessment of the various explanatory factors put forward in the literature reviewed. This assessment was also carried out based on a number of criteria that closely suit our research program. In other words, the assessment component aimed at determining how suitable a source was

to our own objectives, rather than evaluating the validity or impact of the materials surveyed. Table 1 summarizes how we structured the literature review through the use of questions.

By systematizing how we approach the vast literature consulted, we were able to maximize the amount of information derived from each source without losing the overall focus. This was especially relevant given the sheer level of detail in each proposal. At the same time, the system we developed allowed us to draw similarities between the various factors in a straightforward way. While this multiprong approach is perhaps not the most suitable to analyze the fine-grained details in specific change processes, it nevertheless allowed us to draw a broad picture of some of the crucial elements behind linguistic adaptation. In the next section we describe some preliminary results of the ongoing review.

Table 1 - *Topics, search components and questions used to structure the literature review*

Task	*Identification*	*Description*	*Empirical Support*	*Replicability/Specificity*
Information extraction	What name do authors give to this factor?	In which ways does this proposed explanatory factor operate?	What type of data do authors provide to support their proposal?	To which linguistic domain is this proposal applicable? Have authors tested it?
Assessment	To which other proposals can this factor be related?	Are the details of the factor explained in a way that is compatible with language change processes?	How methodologically sound are the data presented? What is the type of sampling technique utilized?	To what extent can this proposal be tested using different data? To what extent can it be used to describe a different linguistic domain?

2.1 Data types

One expected finding given the breadth of our review is that the overlap between the different explanatory factors proposed to explain contact-induced change is rarely explicit. Even when studies tackle similar topics (e.g. 'foreigner-directed speech', 'grammar–based accommodation' and 'audience design', cf. Uther *et al.* 2007; Fehér *et al.* 2019; Arnold *et al.* 2012, among many others), there seems to be little dialogue between studies looking into conceivably related phenomena. This finding may in part result from differences in research

traditions, and/or the existence of area-specific methods. In the above examples, Uther and colleagues observed interactions in lab settings, whereas Fehér and colleagues used an artificial language learning task. Generally speaking, researchers in second language learning tend to observe learners in the context of the classroom, whereas fieldworkers do not typically run perception experiments.

As for the nature of the data, some generalizations emerge from our survey. The first is that many studies are based on in-depth case studies of a single community in a contact situation (e.g. Türker 2000 for Turkish in Norway). Experimental studies, on the other hand, form the basis of the recent second language acquisition literature (e.g. Litcofsky *et al.* 2016), as well as of many bi-/multilingualism studies (e.g. Gyllstad & Wolter 2016 for collocational processing in second-language English speakers). A certain number of studies utilize census-type data, such as population size, number of speakers of a language, data on official languages, and so forth (e.g. Belew & Simpson 2018).

2.2 Types of explanatory factor

Using the procedure we outlined above enabled us to find a number of commonalities between all the different studies investigated nonetheless. We classified the various explanatory factors into four categories: *Cognitive Processes, Interactions between Individuals, Social Networks, and Macro-contexts of Language Use.* These are explained in detail below.

Explanatory factors pertaining to *Cognitive Processes* rely on domain-general processes, such as memory, categorization, perceptual saliency, etc. Proposals classified within this type of explanatory factor tend to focus on the individual as the agent of change, for instance by suggesting that individual learners' inability to hear a phonological distinction may lead to change in the L2 phoneme inventory. The data used in this type of explanatory factor generally come from experiments (e.g. Litcofsky *et al.* 2016), but also from language corpora, and from case studies (e.g. Blevins 2017). Phonological variables tend to predominate in proposals evoking Cognitive Processes. This is the case of Blevins (2017), where 'perceptual magnet' effects are evoked to explain areality in the distribution of sound patterns. Perceptual effects are typically used to account for how phonetic prototypes

function in perception (cf. Kuhl 1991; Kuhl *et al.* 2008). Many of these proposals also apply to other linguistic domains, as in the case of 'metatypy'. Metatypy refers to syntactic and semantic changes that occur in bilingual communities, in which the less dominant language typically changes based on patterns in the more dominant one (Ross 2007: 116).

Explanatory factors classified as *Interactions between Individuals* also tend to focus on domain-general phenomena. One key difference between these factors and Cognitive Processes is that here the focus is on how individual speakers use language to interact with one another. That is, explanations in this category describe change as a product of interactions. For instance, in the case of 'foreigner-directed speech', native speakers would avoid using infrequent or complex constructions when talking to foreigners in an attempt to facilitate communication (see Rothermitch *et al.* 2019 and references therein). The data used in this type of proposal generally derive from case studies (e.g. Ferguson 1975; Berdicevskis 2020), although experiments also figure prominently (e.g. Uther *et al.* 2007; Weatherholtz *et al.* 2014; Chun *et al.* 2016). Studies on adaptive changes resulting from interactions between individuals commonly address morphosyntactic (Weatherholtz *et al.* 2014; Chun *et al.* 2016; Fehér *et al.* 2019) and phonological variables (Uther *et al.* 2007), even though proposals about accommodation in the lexicon have also been made (e.g. Ferguson 1975).

Proposals belonging within the *Social Networks* factor focus on the pathways of information flow between individuals and groups. The investigation of the diffusion of linguistic variables is prominent, and it emphasizes both within-group and between-group communication dynamics. The data used in these studies come from agent-based modeling work (e.g. Fagyal *et al.* 2010; Clem 2016), as well as fieldwork, and from case studies (e.g. Milroy & Milroy 1985; Lippi-Green 1989). The type of data discussed in Social Networks proposals vary widely, depending on the researcher's focus. Social Network proposals seem readily amenable to the study of many types of variables, although replicability may be affected by the highly specific scenarios discussed.

Proposals belonging within *Macro-contexts of Language Use* make reference to the broadest contexts of language use. Such proposals tend to include socio-historical and socioeconomic variables, for instance

when discussing the role of colonization in the adoption of an official language (e.g. Spolsky & Lambert 2005), or the degree of language vitality (Mufwene 2017 and related response articles). Proposals in the Macro-contexts of Language Use often employ census-type data, meaning population size, official language, language of instruction, etc., although some sources also describe case studies. These studies tend to focus on post-hoc analyses, such as in the examination of the impact of ceremonial/religious language on various vernaculars (e.g. Fudge 2005). Because they are so general, Macro-contexts of Language Use explanations lend themselves to the study of most linguistic variables, with lexical variables more readily analyzable.

Uncovering, cataloguing, and grouping the myriad of explanatory factors scattered across related subfields constitutes a first attempt at proposing a framework of linguistic adaptation that better reflects the range of outcomes we encounter as researchers. At present, we are exploring possible overlaps between the four types of explanatory factors, with the ultimate goal of developing a comprehensive model of the mechanisms behind contact-induced change. We remain aware that this future model would still need testing. On the other hand, the information presented in this section has already proven a useful tool in the development of our sociolinguistic questionnaire (Section 5). In the next section we discuss further ways through which our study intends to build bridges across domains, namely by discussing comparative sociolinguistics.

3. *Comparing sociolinguistic scenarios*

In order to research whether linguistic structures adapt to sociolinguistic context across languages, we need to bridge the methodological approaches in language typology and sociolinguistics. A major challenge in this endeavor is the broadly differing methodological traditions in the two disciplines. Our solution is to develop an etic approach to language variation that is based on expert assessments.

Sociolinguistic research tends to focus on naturalistic language data. The data is annotated for language-specific grammatical features and for sociolinguistic features related to individual language users. Other data sources include interviews and questionnaires. The linguistic and social categories used in sociolinguistic research are largely

based on shared norms and have a social reality to the members of the community.

Typological research, on the other hand, focuses on linguistic diversity and largely uses reference grammars as data sources. Comparison across languages is typically based on tools defined by researchers that abstract away from language-particular categories (e.g., Stassen 1985; Haspelmath 2010). The tools used in comparison are thus created by the researcher. Although this approach is the state-of-the-art in language typology, issues about comparability are constantly being discussed (see. e.g. the discussion in special issue 20/2 in the journal *Linguistic Typology*). In typological research linguistic structures are generally analyzed into types that emerge from the variation. Yet, classifying linguistic structures into types is part of the typological method that relies on grammatical descriptions as data sources, rather than an end in itself.

Corpus data have not been as easily available in typology as in sociolinguistics, but the field has been changing recently towards the usage of naturalistic corpus data (e.g., Levshina 2019; Gerdes et al. 2021). This change has been boosted by the greater availability of multilingual parallel texts (e.g. Cysouw & Wälchli 2007) and annotated multilingual corpora, such as Universal Dependencies (Nivre *et al.* 2018). Because we aim at comparing languages across the world, a corpus-driven approach is not feasible for us, owing to the relatively poor availability of systematically annotated corpora especially for most minority languages in the world (but see e.g. the DoReCo project that is beginning to rectify this; Paschen et al. 2020).

In the current project, we develop an approach that bridges sociolinguistic and typological variation by using expert judgments as the basis for comparative data analyses. As mentioned above, the standard data source in typological research is reference grammars. These descriptive works are based on expert judgments about linguistic facts of the language in question. To match this approach in the sociolinguistic part of the project, we rely on expert judgements also for our sociolinguistic data. These data will be collected by using a sociolinguistic questionnaire that will be filled in by field experts collaborating with us. The sociolinguistic questionnaire is described in more detail in Section 5, but we elaborate on some key comparative principles behind it already here.

In our research design, the starting point is the mechanisms of language change, that is, the explanatory factors described in Section 2. We first analyze and classify different explanatory factors into groups and then start formulating broad questions about sociolinguistic factors related to those groups. For instance, related to cognitive processes the questionnaire asks to what extent children are exposed to multilingualism in language acquisition, and related to network structure we ask about the frequency of interaction between groups (e.g. Trudgill 2011).

However, broad questions yield broad answers. We assume that the social sphere is multidimensional and that this needs to be taken seriously in the research design to arrive at more informative answers. In our approach, we try to capture the multidimensionality of social contact by breaking down broad issues into more fine-grained ones. This process results in asking several questions, for instance, related to social networks and not just overall in the community but separately across six predefined social domains. These six social domains (local community, family & kin, social exchange & marriage, trade, labor, and knowledge) are described in Section 5.

We can illustrate the procedure by describing how we operationalize the distinction between 'esoteric' and 'exoteric communication', given that this distinction has been hypothesized to influence language use (Thurston 1987; Wray & Grace 2007; Givón 2009; Trudgill 2011). Esoteric communication takes place among intimates within a small group which means that the interlocutors tend to share much information as well as many norms. Exoteric communication, on the other hand, takes place mostly among strangers within a large group. In this type of communication, there tends to be less shared information between people. The hypothesis is that broad communication types may foster different patterns of interaction and linguistic use, thus ultimately impacting the evolution of different linguistic structures.

The distinction between esoteric and exoteric communication types is captured partly by the labels 'dyad vs. group communication' in our approach. First, we define these distinctions as universally applicable (the numbers given below are based on findings from experimental studies on communication; e.g. Fay *et al.* 2000; Fay & Ellison 2013):

- Dyadic communication refers to communication involving up to four people
- Group-based communication refers to communication involving ten or more people

These definitions are essentially etic and, in this sense, analogous to 'comparative concepts' in language typology (e.g. Haspelmath 2010).

We then formulate our questions so that they yield Likert-scale responses to probe variation a little more carefully than when asking mere categorical questions:

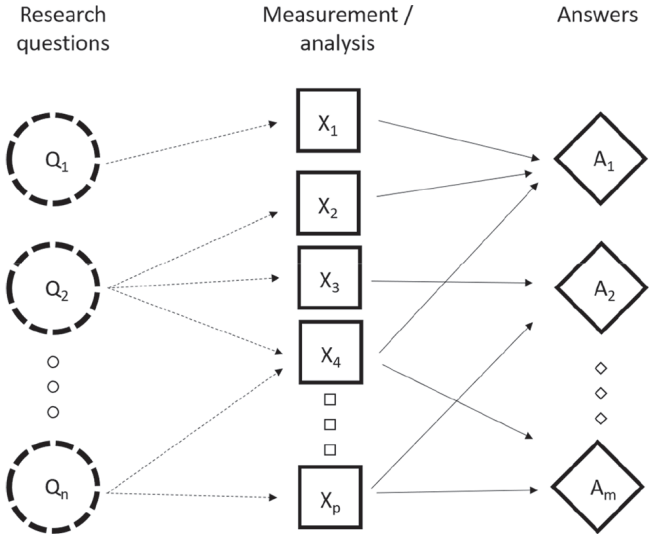
- Question: Are interactions between people in the community typically dyadic or group-based?
- Response options to the question
  - mostly dyadic (~highly esoteric)
  - somewhat dyadic
  - mixed
  - somewhat group-based,
  - mostly group-based (~highly exoteric)

However, this question-response frame by itself only addresses the type of communication in the whole group and thus potentially bypasses group-internal variation. In order to address group-internal variation, we repeat this question across the six predefined social domains. Other questions on social network structure asked in those domains are related to the following:

- Frequency of interaction (strength of connections in social networks)
- Amount of time spent in interaction (strength of connections)
- Effort in reaching other people in the network (strength of connections)
- Type of relationship between the interacting people: friends vs. enemies (nature and strength of connections in social networks)

The approach is schematically summarized in Figure 1.

Figure 1 - *Schematic representation of the research process in typological approach to comparative sociolinguistic research (adapted from Vehkalahti 2019: 122)*



To summarize, the process starts with designing the broad questions. These questions are then operationalized by breaking them down into more fine-grained questions which are used for data collection and analysis. Answers to the broad questions can be arrived at by aggregating the measurements in various ways. First, measurements can be aggregated by sets of questions which are related to the explanatory factors. Second, they can be aggregated by social domain. Third, they can be aggregated in other ways that may be theoretically well-motivated, or overall, to form a bird's-eye-view.

This approach allows us to take seriously the range of sociolinguistic contexts within a community, all the while approaching it from an etic perspective for comparative purposes, as is also done in state-of-the-art approaches to comparison in language typology. In the next section we describe how we sought to include a varied selection of sociolinguistic contexts through our sampling methodology.

4. *Sampling in sociolinguistic typology*

Investigating linguistic adaptation from a typological perspective while comparing sociolinguistic scenarios requires a diverse sample of speech communities from around the world. The sampling methodology developed within the GramAdapt project departs from existing sampling methods in language typology (for overviews, see Bakker 2011 and Miestamo *et al.* 2016.) in that we use sets of languages in contact, rather than individual languages, as the sampling unit. By definition, language contact involves two or more communities in interaction. Thus, understanding contact interactions and their repercussions on language structure requires that the unit of comparison include more than one language.

Our project uses established conventions in language contact research to define the internal structuring of our multi-language sampling units. The contact literature frames contact phenomena as pairwise interactions. In each of these pairs, one language counts as the potential recipient, and the other as the potential source of contact effects (Winford 2010: 171). We adopt this pairwise representation of contact scenarios and define the two languages in each pair as the *Focus Language* and *Neighbor Language,* respectively.

The *Focus Language* is the language whose potential linguistic adaptations we study. The *Neighbor Language* is a language identified in the reference materials to be in contact with the Focus Language. Specifically, the project investigates the Neighbor language's influence on the Focus Language by considering both patterns of language structures and social contact. Our sampling units also include a third language, termed the Benchmark Language, which allows us to disentangle areal diffusion from inheritance in the Focus Language. The Benchmark Language is genealogically related to the Focus Language, but is not in contact with either the Focus or the Neighbor Language. As such, the Benchmark Language serves as a parameter against which to test the impact of contact on the Focus language. Examples are given in Table 2.

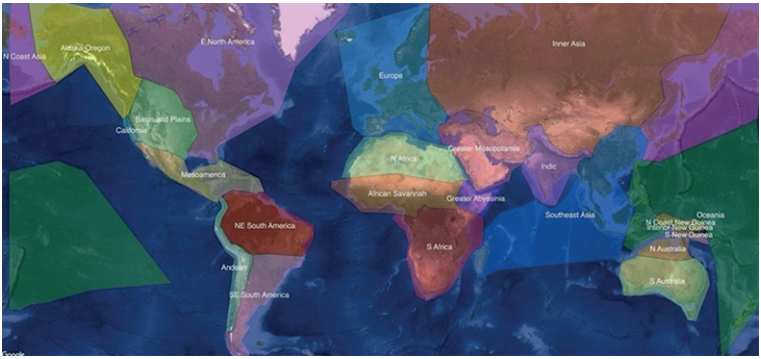
We are aware that the analysis of contact situations through the sets we propose has limitations. For instance, contact influence on the Focus Languages may come from other languages than those selected as Neighbor Languages in a given set. Moreover, the Neighbor Languages themselves may be affected by contact with the Focus Languages (and others). However, given that language contact minimally occurs between users of two languages, limiting our perspective to contact influences on

the Focus Language provides us with a unit of comparison that is schematic enough to be in principle applicable to any contact scenario worldwide. Similarly, restricting the choice of Benchmark to only one language per contact set only offers a partial view of the diachronic processes behind the retentions or innovations of linguistic features in the Focus Language. Nevertheless, sampling one Benchmark per contact set among the Focus Language's own relatives provides us with some measure of comparison that is external to the contact situation at stake, while at the same time controlling for genealogical relatedness.

To summarize, the proposed sampling technique establishes a principled way of capturing contact scenarios for the purpose of worldwide typological comparisons. We argue here that this is a crucial first step towards large scale studies of linguistic adaptation.[2]

Four sets of external criteria guided the compilation of the sample: (1) geographical area, (2) independently reported contact scenarios, (3) genealogical distance between languages in contact, and (4) availability of living experts to collaborate with. These four criteria allowed us to build a typological dataset that is geographically and genealogically stratified, in addition to being independent of our own assessment of a given contact situation. Each criterion is explained in more detail below.

Figure 2 - *The 24 Autotyp areas (Bickel et al. 2017, used under CC-BY 4.0 license)*



[2] In a way, all sampling methods in language typology involve some degree of coarseness in the framing of the units of comparison. For instance, probability samples, which are used to test statistical tendencies in the worldwide distribution of language structures, are typically constructed by extracting one (or a few) language(s) per genealogical unit. This tends to restrict the representation of the range of linguistic diversity attested in a given language family to the data point(s) chosen for that particular family.

The first criterion, geographical area, uses the 24 geographical areas established in the Autotyp database (Nichols *et al.* 2013; Bickel *et al.* 2017). The Autotyp areas (see Figure 2) derive from archeological, anthropological, historical, and genetic data, and are thus established independently of linguistic features (Nichols *et al.* 2013: 6). As such, using the Autotyp areas allows us to randomize language selection in a way that is blind to the goals of our own study. The sample currently consists of two sets of three languages for most Autotyp areas, totaling 150 languages. Two areas, Northeast South America, and Southern New Guinea, each provide three sets given the high degree of linguistic diversity found in South America and New Guinea as a whole (e.g. Dahl 2008; Hammarström 2016). The three language sets sampled for Southern New Guinea are shown in Table 2.

Table 2 - *Language sets sampled for the Trans Fly contact zone in Southern New Guinea (ISO 639-3 codes and language families are shown in parentheses)*

Autotyp Area (Source)		*Focus*	*Neighbor*	*Benchmark*
Southern New Guinea Evans *et al. (2018)*	*SET 1*[3]	Nen (nqn; Yam)	Idi (Idi; Pahoturi River)	Yei (jei; Yam)
	SET 2	Coastal Marind (mrz; Anim)	Marori (mok; Marori)	Warkay-Bipim (bgv; Anim)
	SET 3	Koiari (kbk; Kolarian)	Motu (meu; Austronesian)	Ese (mcq; Kolarian)

The second criterion, namely independently reported contact scenarios, guides the choice of which languages to analyze from each area. This criterion reflects descriptions of language contact in the literature by experts in given areas of the world, such as in macro-area surveys of contact situations and/or areal linguistics. For instance, the surveys of languages of Southern New Guinea by Evans (2012) and Evans *et al.* (2018) describe several cases of contact situations in the area, from which we drew our own sets.

[3] As with all other sampled sets, we use Glottolog as a reference system for the genealogical classification of the languages of this set. However, we remain aware that, in this particular case, Glottolog's internal groupings for the Yam family do not completely overlap with state-of-the-art comparative reconstructions (cf. Evans et al. 2018: 68)

The choice of the three languages also rests on the third criterion behind our sampling procedure: genealogical distance. Each set in our sample includes only Focus and Neighbor Languages that belong to distinct language families as classified in Glottolog (Hammarström *et al.* 2020). The choice of genealogically unrelated languages helps ensure that linguistic effects of the Neighbor Language on the Focus Language stem from the contact situation. On the other hand, the Benchmark Language is chosen among the Focus Language's relatives, with the added condition that it must not be part of the Focus/Neighbor contact zone.[4] The degree of relatedness between Focus and Benchmark languages varies across the sampled sets, depending on the genealogy and contact history of individual language families. For instance, considering set 3 in Table 2, according to Glottolog's classification, Koiari and Ese, the Focus and Benchmark languages, belong to two distinct upper level subdivisions of the Koiaran family. Koiari is part of the Koiaric subgrouping while Ese is part of the Baraic subgrouping and is spoken outside the contact zone where the Focus and Neighbor language are found.[5]

The fourth and final selection criterion is the availability of experts who could describe the contact scenario between Focus and Neighbor Language. This criterion follows from the fact that collaboration with experts is at the core of the data collection through our sociolinguistic questionnaire (Section 5). When selecting our pool of experts, we also strived to involve researchers from the communities under study, in an attempt to engage with different academic cultures and more local perspectives on the contact scenarios at stake.

Following the procedure outlined above, we identified 50 contact sets from all parts of the world and across all Autotyp Areas.[6] The sample currently includes 50 Focus Languages distributed across 36 lan-

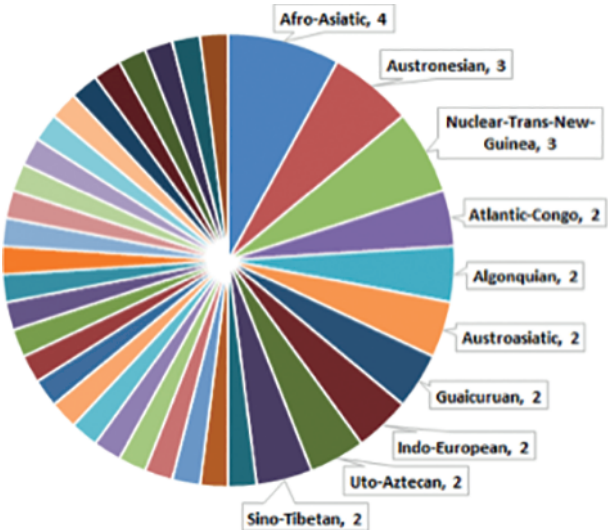
[4] One consequence of this setup is that while Neighbor Languages can be isolates (e.g. Basque in the Gascon-Basque contact pair), Focus Languages cannot. By establishing that Focus Languages should be compared with Benchmark Languages, isolates are by definition excluded.

[5] We will apply the Diversity Value Method (Rijkhoff et al. 1993; Rijkhoff & Bakker 1998) on all pairs of Focus and Benchmark languages as a measure of genealogical relatedness. This relatedness score will be used as one of the independent variables to factor into the analyses of the linguistic and sociolinguistic data.

[6] Sample size was set at 50 sets (i.e. 150 languages in total) so as to maximize diversity in sample composition, while keeping a manageable workload given the time frame of the project.

guage families (as per Glottolog's classification), with an average of 1.4 languages per family.[7] While languages from large and geographically widespread language families (e.g. Afro-Asiatic, Austronesian, Nuclear-Trans-New-Guinea) are represented more than once in the sample in different roles (i.e. as Focus, Neighbor, or Benchmark languages), the majority of the sampled language families are small. Therefore, most language families are represented only once in our dataset. Figure 4 illustrates our sample by reporting the distribution of Focus Languages per language family. The figure also details the number of Focus Languages selected for the ten language families with more than one representative included in the sample.

Figure 3 - *Distribution of Focus Languages per language family, detailing the number of Focus languages from families with more than just one representative in our sample. Focus Languages Genealogical Affiliation (N=50)*



The selection process was fully independent of the linguistic variables that we use to assess the linguistic outcomes of a given contact situation (Section 6). The language selection process was also independent of the

[7] We only report on the Focus Languages here because the choice of Neighbor Languages is still under discussion for a few sets. As mentioned above, Benchmark languages belong to the same families as Focus languages.

social history of a given contact situation, which is what we aim to study in depth through the sociolinguistic questionnaire, explained in the next Section. To summarize, the sampling technique presented here is, to the best of our knowledge, the first attempt to develop a typological dataset that is specifically designed for worldwide studies of linguistic adaptation. As we hope to have shown, the methodology is transparent and replicable, and has the potential to provide a tool for investigating the influence of sociolinguistic environments on the distribution of linguistic diversity.

5. *Sociolinguistic questionnaire*

As stated in Section 3, this project attempts to bridge sociolinguistic and typological variation by using expert judgments as the basis for comparative data analyses. The GramAdapt questionnaire is a practical manifestation of this attempt.

The questionnaire comprises two parts. The first part is the Overview Questionnaire, which is designed to match prior studies investigating correlations between linguistic and macro socio-cultural factors. These macro-factors include speaker population (Lupyan & Dale 2010), degrees of political complexity (Currie & Mace 2009), and types of marriage systems (Bowern 2010 in hunter-gatherer societies). The second part is the Domains Questionnaire, which will be the focus for the remainder of this paper.

The Domains Questionnaire design is based on the well-established finding that different kinds of interactional situations beget different kinds of linguistic behaviors; known in the literature as "the domains of language use" (Fishman 1965). We have identified six social domains that typically have norms and modes of interaction particular to that domain, and are often attested as domains of social contact.

The six social domains are: local community, family & kin, social exchange & marriage, trade, labor, and knowledge. We developed operational definitions for each domain with the aim to make them as cross-culturally applicable as possible, which will also aid in the development of comparative tools as outlined in Section 3. Given the range of societal diversity we see across contact situations in space and time, the domains must be flexible enough to capture the dynamics of small and egalitarian communities, right through to communities that are part of larger, socially stratified and administratively complex societies.

While there are inevitable overlaps between the domains when considering real communities, we have attempted to create an operational definition for each domain, which is as outlined below:

- Local Community: Concerning interactions beyond kin, and outside institutionalized practices related to trade, labor, and knowledge. For example, a village, a band, or a neighborhood in a town.
- Family & Kin: Concerning the interaction between members of family and kin. This domain typically encompasses child bearing and rearing, as well as food production and consumption.
- Social Exchange & Marriage: Concerning practices of exchange which regulate relationships between individuals and groups, within and across societies. This domain encompasses practices of gift and ceremonial exchange, as well as marriage exchange.
- Trade: Concerning transactions of objects and services. The prototype is a transaction of commodities. The mode of transaction can be monetary, barter-based, etc.
- Labor: Concerning practices and relationships that revolve around economic activity and production.
- Knowledge: Concerning knowledge transfer that is structured in culturally specific ways. This domain prototypically covers practices that revolve around education and religion.

As stated earlier, the goal of the Domains Questionnaire is to seek data on contact dynamics in each of the six domains along the four explanatory factors for contact-induced change illustrated in Section 2, that is cognitive processes, interaction between individuals, social networks, and macro-contexts of language use (see Section 2 and 3 for illustrations of each of these factors and their relation to our comparative sociolinguistic approach). The data will be collected for those domains where social contact occurs between Focus and Neighbor groups. The response across the social domains would then scale up to a general profile of social contact, from the bottom-up.

To give an illustration, for the explanatory factor Social Networks, we have seven questions designed to get a sense of social network scope. For example, one question asks about the frequency of interaction between Focus and Neighbor Group peoples (question S1 in table 4). Another asks about interaction type: whether it

is dyadic, or group-based and broadcast-like (Section 3). For every domain where social contact occurs, the respondent will provide an answer to these questions.

If, for a particular subset of the Domains Questionnaire there is no contact between the groups, the respondent will simply skip that particular domain. Once the questionnaire is filled, we will have an aggregate view of how dense contact is across domains. We may surmise that the more social domains where Focus and Neighbor Group peoples interact, the denser the contact.

We illustrate how the Questionnaire is structured with examples of two situations: one in New Guinea and the other in India. In one situation, we have social contact between speakers of Nen (Yam), and Idi (Pahoturi River) in southern Papua New Guinea. In the other situation, we have social contact between speakers of Marathi (Indo-Aryan) and Kannada (Dravidian) in Kupwar, India. The socio-cultural configuration of the two situations is distinct. In the New Guinea case, the groups are non-hierarchical, and based on subsistence horticulture. In India, we have a caste-based stratified society which is part of a larger socio-political complex. In New Guinea, the total number of speakers is quite small: 350 for Nen, and 800 for Idi (Evans *et al.* 2018: 645-646). The total number of Marathi and Kannada speakers in Maharashtra state is 77.5 million and 1 million respectively (Office of the Registrar General & Census Commissioners 2011). More details on the socio-cultural configurations of the two situations is presented in Table 3.

Table 3 - *Overview of some societal-demographic characteristics of Nen/Idi contact pair, and Marathi/Kannada contact pair*

Contact Pair	Focus Group = Nen (Yam); Neighbor Group = Idi (Pahoturi River)	Focus Group = Marathi (Indo-Aryan); Neighbor Group= Kannada (Dravidian)
Autotyp Area	Southern New Guinea	Indic
Language and social affiliation	Äkämar tribe people claim Nen as their language. Gunduma people claim Idi as their language	Low-caste Hindus speak Marathi, land owning Jains and Lingayat crafts people speak Kannada

Socioeconomic hierarchies and subsistence pattern	Non-hierarchical, subsistence horticulture	Caste-based professional hierarchy, complex agriculture
Relationship to larger socio-political structures	Historically not part of a centralized state until effective colonization in the 1960s, currently national infrastructure is mostly absent	Historically parts of various empires, currently part of Maharashtra state of the Republic of India with Marathi the official language of education and administration in Maharashtra state
Language ideologies	Egalitarian multilingualism (as defined by François 2012) with a relative absence of major languages (i.e. Tok Pisin)	Caste and religion-based linguistic divisions
Speaker population	Approximate total speaker population: Nen = 350 Idi = 800	Approximate total speaker population: Marathi = 77.5 million in Maharashtra state; 83 million in the whole of India Kannada = 1 million in Maharashtra state; 43.5 million in the whole of India

Let us consider three social domains: Local Community, Family & Kin, and Trade. Under the operational definitions we produced for the questionnaire, the New Guinea situation shows social contact in two domains: Local Community, and Family & Kin. In our definition of trade, we emphasize the fact that a given transaction presupposes an expectation of immediate or future return. This configuration does not apply to the New Guinea scenario.[8] The respondent will therefore answer the questions for Local Community and Family & Kin, but not for Trade. The India situation, on the other hand, shows social contact in Local Community and Trade, but not Family & Kin. Language is a property of caste membership (Gumperz & Wilson 1971; Kulkarni-Joshi 2016), and people of different castes do not intermarry. The respondent will therefore answer the questionnaire for the domains of local community and trade, but not family & kin.

Table 4 provides an example of responses for the Nen/Idi and Marathi/Kupwar contact situations in the domain of Local

[8] The social domain concerned with what we call "social exchange" would cover situations like the New Guinea case where relationship building is either an explicit or crucial motivator of conducting exchange.

Community. The possible responses for the Social Network set of questions are based on a five-point scale,[9] as shown earlier on in Section 3. Eri Kashima provided the responses for the purpose of this demonstration. She is an expert of Nmbo (ncm; Yam) and has worked with Nen speakers in the context of studying the sociolinguistics of the area. Eri Kashima also filled out the Marathi/Kannada response for the purposes of this demonstration, based on Gumperz & Wilson (1971) and Kulkarni-Joshi (2015; 2016). The final version of the questionnaire will be filled out by an expert who has worked in the Marathi/Kannada context. The sample answers for Local Community indicate that the Nen/Idi contact situation may be denser than Marathi/Kannada. These responses will form one portion of our bottom-up characterization of social contact in these situations, which can then be compared across the sample set.

One major challenge in designing the questionnaire has been setting the time-frame of the questionnaire response. In order to test for linguistic adaptation, one would need to know about the contact situation before present; since linguistic changes visible at the present would have adapted to a given socio-historical context in the past. Given the challenges in generalizing across the time-depth of contact scenarios worldwide, we refrain from establishing any a priori chronological cut-off points for the contact situations in this questionnaire. Instead, we ask respondents to (a) assess the duration of contact between Focus and Neighbor Group in a given social domain, and (b) identify the time frame of densest contact between Focus and Neighbor Group in said social domain. The questions should be answered from the perspective of this time frame. In doing so, we hope to gain an understanding of the contact situations at stake that is maximally entrenched in the specifics of their linguistic and social ecologies. There are multiple ways in which this time-frame issue could be dealt with, but our solution strikes a balance between the availability of data and the diversity of contact situations from a global perspective.

[9] Not all of the questions, however, are based on a five-point scale. There are many questions with binary "yes/no" answers.

Table 4 - *An example of answers to the Social Network questions (S-Set). These sets of questions require an answer on a five-point scale. The ID number of questions (S1, S3 etc.) is non-contiguous as question S2 was retired during the design process. The expression of the question in this demonstration shows the essence of the question, rather than the final formulation in the questionnaire*

Network Structure Question	*Nen & Idi (New Guinea)*	*Marathi & Kannada (India)*
S1: How often do Focus Group people typically interact with Neighbor Group people?	3	4
S3: How many people are typically involved in interactions between Focus Group and Neighbor Group? Is it more-or-less dyadic or group-based?	4	2
S4: How physically proximate to each other are Focus Group people and Neighbor Group people in this domain?	4	4
S5: How friendly are Focus Group people and Neighbor Group people in this domain?	5	3
S6: What is the proportion of total Focus Group people who have opportunities for contact with Neighbor Group people	5	2
S7: What is the proportion of total Neighbor Group people who have opportunities for contact with Focus Group people	3	2
Network Density Score for Local Community	24	17

6. *Linguistic variables*

Aside from collecting data on social contact between Focus and Neighbor Groups as detailed in Section 5, we will also work with a selection of linguistic variables from different domains of language structure ranging from phonology, morphosyntax, and the lexicon. The coding for each of these variables will be based on established literature on relevant patterns of crosslinguistic variation and diachronic change in these domains. The variables will be used to test hypotheses about linguistic adaptation in the languages of the sample; that is, to investigate whether any structural features of the Focus Languages may be attributed to contact with the respective Neighbor languages or whether other processes (i.e. plain retention and/or language internal evolution) may be at stake. The linguistic variables are chosen among renown cross-linguistically common structures in the domain

of phonology, morphosyntax, and the lexicon. While the variable selection process is still ongoing, we have a preliminary plan of action for the domains that we would like to consider.

Within phonology, we will be looking at suprasegmental patterns, such as syllable structure and word stress patterns. As recently demonstrated in Napoleão de Souza & Sinnemäki (revised), suprasegmental features represent a promising testing ground to investigate phonological change in contact situations. Within morphosyntax, we will be focusing on nominal number and locus of marking. Number is the most frequent nominal category cross-linguistically (Corbett 2000) and is also relatively well studied from a typological and language contact perspective, which facilitates both variable selection and data collection using reference materials (for a discussion of number marking and language contact dynamics see, for instance, Roberts & Bresnan 2008 and Igartua 2015). The same could be said about locus of marking (Nichols 1992). Locus of marking concerns syntactic relations and how they are morphologically marked within phrases and the clause. Syntactic relations are often morphologically marked in languages either on the head or the dependent of the construction, and these markings are prone to change especially under heavy language contact (Roberts & Bresnan 2008). Finally, in the domain of the lexicon, we will study demonstrative systems. Our focus will be on adnominal demonstratives. These are often mentioned in the contact linguistics literature, but have never been systematically studied from the perspective of language adaptation, which we will be testing for the first time.

Working on a selection of linguistic features from a diverse range of structural domains increases the chance that at least some of the variables of choice will be meaningful for the purpose of identifying and describing processes of adaptation. Moreover, this procedure allows us to test whether the likelihood of adaptation differs across domains – e.g. if prosodic features turn out to be more likely to undergo adaptive changes than demonstrative systems or number systems. Finally, the method enables us to investigate whether the types of attested language changes, as well as their frequency of occurrence, vary depending on the length and intensity of contact, which we estimate through the sociolinguistic questionnaire.

All sampled languages will be coded for the same set of linguistic features. This procedure allows us to run comparable analyses across sampled languages and contact sets. Through these comparisons, we can then assess the possibility that some variables may be more relevant to certain contact sets than others. We are aware that linguistic adaptations may occur elsewhere than in those domains we have chosen to investigate.

A crucial issue in the linguistic data collection process concerns variable design. Much of the research on linguistic adaptation has focused on number of distinctions, for instance, in phoneme inventory sizes (Hay & Bauer 2007), number of cases (Bentz & Winter 2013; Sinnemäki 2020), and number of gender distinctions (Sinnemäki & Di Garbo 2018; Dahl 2019). However, as suggested in recent research (cf. Sinnemäki & Di Garbo 2018; Verkerk & Di Garbo accepted, 2021, regarding grammatical gender), counting the number of distinctions in a grammatical domain may not always be the best way to assess patterns of linguistic adaptation.

An alternative to the number-of-distinctions approach is to consider the processes of restructuring that language structures undergo from a cross-linguistic perspective, and how they may correlate with different types of sociolinguistic scenarios. We embrace this alternative approach, which has also been validated in recent research on linguistic adaptation and contact-induced change.

With respect to linguistic adaptation in morphosyntax, researchers have shown that contact situations characterized by high proportions of adult L2 learning favor processes of restructuring that increase the transparency and compositionality of morphosyntactic paradigms (Kusters 2003; Trudgill 2011; Kempe & Brooks 2018). For instance, patterns of grammatical gender marking only marginally shaped by semantic criteria may become fundamentally restructured around the encoding of animacy distinctions. By studying the evolution of gender systems in northwestern Bantu languages, Di Garbo & Verkerk (accepted, 2021) and Verkerk & Di Garbo (accepted, 2021) find that these highly transparent gender systems abound in languages with a history of intense language contact and/or language shift[10].

[10] Previous studies that looked at the sociolinguistic typology of grammatical gender by using the number of gender distinctions as the linguistic variable of interest were not able to show any such effect (Sinnemäki & Di Garbo 2018; Dahl 2019).

Similarly, Napoleão de Souza & Sinnemäki (revised) demonstrate that looking at processes of restructuring in suprasegmental features may be more informative than a simple assessment of the presence vs. absence of some phonological variables. The authors claim that focusing on processes may advance our understanding of the impact of language contact on phonological structure.

In view of the evidence presented above, for each of the linguistic variables of choice, we will develop our coding design in a way that is informed by research in typology, historical linguistics, sociolinguistics, and studies of bi-/multilingual language use.

7. *Concluding remarks*

In this paper we outlined the approach developed within the GramAdapt project with the aim of establishing appropriate concepts and methods for investigating the relationship between languages and their sociolinguistic environments. The proposed framework can be used to study sociolinguistic correlates of linguistic diversity and language change in three ways: (1) through the analysis of causal factors related to language change, (2) through a novel sampling technique simultaneously addressing selection of communities, sociolinguistic features, and linguistic features, and (3) through generalizations from empirically-grounded cross-cultural and cross-linguistic comparisons. It is our hope that this approach to worldwide comparisons of language structures and communities will set a new ground for the typological study of linguistic adaptation.

Acknowledgements

This research has received funding by the European Research Council (ERC), grant no 805371 to Kaius Sinnemäki (PI). We wish to acknowledge the audience of the SLI workshop on "Sociolinguistics and linguistic typology: towards an integrated approach to the study of linguistic variation" (September 2020) and, in particular, the workshop organizers and editors of this volume Silvia Ballarè and Guglielmo Inglese. For feedback on the research design, we are especially grateful to Maria Khachaturyan, Olesya Khanina, Friederike Lüpke, Susanne Michaelis, Brigitte Pakendorf, Ksenia Shagal, and

Max Wahlström, as well as to the audience of the GramAdapt online seminar series (November – December 2020). For help and support with the manuscript layout, we thank Janne Loisa. The usual disclaimers apply.

Bibliography

Atkinson, Quentin. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332(6027). 346-349.

Arnold, Jennifer E., M. Kahn, Jason & C. Pancani, Giulia. 2012. Audience design affects acoustic reduction via production facilitation. *Psychonomic bulletin & review* 19(3). 505-512.

Bakker, Dik. 2011. Language Sampling. In Song, Jae Jung (ed.) *Handbook of Linguistic Typology*, 100-127. Oxford: Oxford University Press.

Belew, Anna & Simpson, Sean. 2018. The status of the world's endangered languages. In Rehg, Kenneth L. & Campbell, Lyle (eds.) *Oxford handbook of endangered languages*, 21-47. Oxford: Oxford University Press.

Bentz, Christian & Winter, Bodo. 2013. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change* 3. 1-27.

Berdicevskis, Aleksandrs. 2020. Foreigner-directed speech is simpler than native-directed: Evidence from social media. *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*. 163-172.

Bickel, Balthasar, Nichols, Johanna, Zakharko, Taras, Witzlack-Makarevich, Alena, Hildebrandt, Kristine, Rießler, Michael, Bierkandt, Lennart, Zúñiga, Fernando & Lowe, John B. 2017. *The AUTOTYP typological databases*. Version 0.1.0 https://github.com/autotyp/autotyp-data/tree/0.1.0

Blasi, Damián E., Moran, Steven, Moisik, Scott R., Widmer, Paul, Dediu, Dan & Bickel, Balthasar. 2019. Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363(6432). doi: 10.1126/science.aav3218

Blevins, Juliette. 2017. Areal Sound Patterns: From Perceptual Magnets to Stone Soup. In Hickey, Raymond (ed.), *The Cambridge Handbook of Areal Linguistics*. 88-121. Cambridge: Cambridge University Press.

Bowern, Claire. 2010. Correlates of Language Change in Hunter-Gatherer and Other 'Small' Languages. *Language and Linguistics Compass* 4(8). 665-679.

Chun, Eunjin, Barrow, Julia, & Kaan, Edith. 2016. Native English speakers' structural alignment mediated by foreign-accented speech. *Linguistics Vanguard* 2(s1). 1-10.

Clem, Emily. 2016. Social network structure, accommodation, and language change. *UC Berkeley Phonetics and Phonology Lab Annual Report*, 83-102.

Corbett, Greville G. 2000. *Number*. Cambridge: Cambridge University Press.

Currie, Thomas E. & Mace, Ruth. 2009. Political complexity predicts the spread of ethnolinguistic groups. *Proceedings of the National Academy of Sciences* 106(18). 7339-7344.

Cysouw, Michael & Wälchli, Bernhard. 2007. Parallel texts: using translational equivalents in linguistic typology. *STUF – Language Typology and Universals* 60(2). 95-99. https://doi.org/10.1524/stuf.2007.60.2.95.

Dahl, Östen. 2004. *The growth and maintenance of complexity.* Amsterdam: John Benjamins.

Dahl, Östen. 2008. An exercise in a posteriori language sampling. *STUF – Language Typology and Universals* 61(3). 208-220.

Dahl, Östen. 2019. Gender: exoteric or esoteric? In Di Garbo, Francesca, Olsson, Bruno & Wälchli, Bernhard (eds.), *Grammatical gender and linguistic complexity, vol. I: General issues and specific studies*, 53-61. Berlin: Language Science Press.

Di Garbo, Francesca & Verkerk, Annemarie. Accepted, 2021. *A typology of northwestern Bantu gender systems*. To appear in *Linguistics*.

Evans, Nicholas. 2012. Even more diverse than we thought: the multiplicity of Trans-Fly languages. In Evans, Nicholas and Klamer, Marian (eds.) *Melanesian Languages on the Edge of Asia: Challenges for the 21st Century*, 109-149. Honolulu: University of Hawai'i Press.

Evans, Nicholas, Arka, I. Wayan, Carroll, Matthew J., Choi, Yun Jung, Döhler, Christian, Gast, Volker, Kashima, Eri, Mittag, Emile, Olsson, Bruno, Quinn, Kyla, Schokkin, Dineke, Tama, Phillip, van Tongeren, Charlotte, & Siegel, Jeff. 2018. The Languages of Southern New Guinea. In Palmer, Bill (ed.), *The Languages and Linguistics of New Guinea: A Comprehensive Guide*, 640-774. Berlin: De Gruyter Mouton.

Everett, Caleb, Blasi, Damián E. & Roberts, Seán G. 2015. Climate, vocal folds, and tonal languages: connecting the physiological and geographic dots. *PNAS* 112. 1322-1327.

Fagyal, Zsuzanna., Swarup, Samarth, Escobar, Anna María, Gasser, Les, & Lakkaraju, Kiran. 2010. Centers, Peripheries, and Popularity: The Emergence of Norms in Simulated Networks of Linguistic Influence. *University of Pennsylvania Working Papers in Linguistics* 15(2). 81-90.

Fay, Nicholas, & Ellison, T. Mark. 2013. The Cultural Evolution of Human Communication Systems in Different Sized Populations: Usability Trumps Learnability. *PLoS One* 8(8). https://doi.org/10.1371/journal.pone.0071781

Fay, Nicholas, Garrod, Simon & Carletta, Jean. 2000. Group discussion as interactive dialogue of group monologue: The Influence of Group Size. *Psychological Science* 11(6). 481-486.

Fehér, Olga, Ritt, Nikolaus & Smith, Kenny. 2019. Asymmetric accommodation during interaction leads to the regularisation of linguistic variants. *Journal of Memory and Language* 109. 104036.

Ferguson, Charles A. 1975. Toward a characterization of English foreigner talk. *Anthropological Linguistics* 17. 1-14.

Fishman, Joshua A. 1965. Who Speaks What Language to Whom and When? *La Linguistique* 1(2). 67-88.

Fudge, Erik. 2005. Religion and language. In Keith Brown (ed.) *Encyclopedia of language and linguistics*. Elsevier.

Gerdes, Kim, Kahane, Sylvain & Chen, Xinying. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa: A Journal of General Linguistics* 6(1). 17. http://doi.org/10.5334/gjgl.764

Givón, Talmy 2009. *Syntactic Complexity: Diachrony, Acquisition, Neurology, Evolution.* Amsterdam: John Benjamins.

Gumperz, John, & Wilson, Robert. 1971. Convergence and creolization: a case from the Indo-Aryan/Dravidian border in India. In Hymes, Dell H. (ed.), *Pidginization and Creolization of Languages*, 153-169. Cambridge: Cambridge University Press.

Gyllstad, Henrik & Wolter, Brent. 2016. Collocational processing in light of the phraseological continuum model: Does semantic transparency matter? *Language Learning* 66(2). 296-323.

Hammarström, Harald. 2016. Linguistic diversity and language evolution. *Journal of Language Evolution* 1(1). 19-29. https://doi.org/10.1093/jole/lzw002

Hammarström, Harald, Forkel, Robert, Haspelmath, Martin & Bank, Sebastian. 2020. *Glottolog database 4.3*. Jena: Max Planck Institute for the Science of Human History. doi:10.5281/zenodo.3754591.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663-687. http://www.jstor.org/stable/40961695.

Hay, Jennifer & Bauer, Laurie. 2007. Phoneme inventory size and population size. *Language 83*. 388-400.

Hruschka, Daniel J., Christiansen, Morten H., Blythe, Richard A., Croft, William, Heggarty, Paul, Mufwene, Salikoko S., Pierrehumbert, Janet B., & Poplack, Shana. 2009. Building social cognitive models of language change. *Trends in Cognitive Sciences*, 13(11). 464-469. https://doi.org/10.1016/j.tics.2009.08.008

Igartua, Ivan. 2015. From cumulative to separative exponence: Reversing the morphological cycle. *Language* 91(3). 676-722.

Kempe, Vera & Brooks, Patricia J. 2018. Linking adult second language learning and diachronic change: A cautionary note. *Frontiers in Psychology 9*. doi:10.3389/fpsyg.2018.00480. https://www.frontiersin.org/article/10.3389/fpsyg.2018.00480.

Kousters, Wouter. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflections*. Utrecht: LOT Dissertation Series.

Kuhl, Patricia. 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories. Monkeys do not. *Perception & psychophysics* 50. 93-107. 10.3758/BF03212211.

Kuhl, Patricia K., Conboy, Barbara T., Coffey-Corina, Sharon, Padden, Denise, Rivera-Gaxiola, Maritza, & Nelson, Tobey. 2008. Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1493). 979-1000.

Kulkarni-Joshi, Sonal. 2015. Religion and language variation in a convergence area: The view from the border town of Kupwar post-linguistic reorganisation of Indian states. *Language and Communication* 42. 75-85.

Kulkarni-Joshi, Sonal. 2016. Forty years of language contact and change in Kupwar: A critical assessment of the intertranslatability model. *Journal of South Asian Languages and Linguistics* 3(2). 147-174.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533-572.

Lippi-Green, Rosina L. 1989. Social network integration and language change in progress in a rural alpine village. *Language in Society* 18(2). 213-234.

Litcofsky, Kaitlyn A., Tanner, Darren & van Hell, Janet G. 2016. Effects of language experience, use, and cognitive functioning on bilingual word production and comprehension. *International Journal of Bilingualism* 20(6). 666-683.

Lupyan, Gary & Dale, Rick. 2010. Language structure is partly determined by social structure. *PLOS One* 5(1). 1-10.

Lupyan, Gary & Dale, Rick. 2016. Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity. *Trends in Cognitive Sciences* 20(1). 649-660.

Miestamo, Matti, Bakker, Dik & Arppe, Antti. 2016. Sampling for variety. *Linguistic Typology* 20(2). doi:10.1515/lingty-2016-0006.

Milroy, James & Milroy, Lesley. 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics* 21(2). 339-384.

Mufwene, Salikoko S. 2017. Language vitality: The weak theoretical underpinning of what can be an exciting research area. *Language* 93(4). 202– 223.

Napoleão de Souza, Ricardo & Sinnemäki, Kaius. Revised. *Beyond segment inventories: Phonological complexity measures and suprasegmental variables in contact situations.*

Nichols, Johanna. 1992. *Linguistic Diversity in Space and Time*. Chicago, IL: University of Chicago Press.

Nichols, Johanna, Witzlack-Makarevich, Alena & Bickel, Balthasar. 2013. *The AUTOTYP genealogy and geography database: 2013 release*. http://www.spw.uzh.ch/autotyp/.

Nivre, Joakim, Abrams, Mitchell, Agić, Željko, et al. 2018. *Universal Dependencies 2.3. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University*. http://hdl.handle.net/11234/1-2895.

Office of the Registrar General & Census Commissioner. 2011. *Part A: Distribution of the 22 scheduled languages-India/States/Union Territories*. Retrieved from [https://censusindia.gov.in/2011Census/Language_MTs.html]

Paschen, Ludger, Delafontaine, François, Draxler, Christoph, Fuchs, Susanne, Stave, Matthew & Seifart, Frank. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). *Proceedings of the 12th Conference on Language Resources and*

Evaluation (LREC 2020), 2657-2666. Marseille: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.324.pdf

Rijkhoff, Jan & Bakker, Dik. 1998. Language sampling. *Linguistic Typology* 2-3(3). 263-314.

Rijkhoff, Jan, Bakker, Dik, Hengeveld, Kees & Kahrel, Peter. 1993. A method of language sampling. *Studies in Language* 17. 169-203.

Roberts, Sarah and Joan Bresnan. 2008. Retained inflectional morphology in pidgins: A typological study. *Linguistic Typology* 12. 269-302.

Ross, Malcom. 2007. Calquing and metatypy. *Journal of Language Contact* 1(1). 116-143.

Rothermich, Kathrin, Harris, Havan L., Sewell, Kerry &. Bobb, Susan C. 2019. Listener impressions of foreigner-directed speech: A systematic review. *Speech Communication* 112. 22-29.

Sinnemäki, Kaius. 2020. Linguistic system and sociolinguistic environment as competing factors in linguistic variation: A typological approach. *Journal of Historical Sociolinguistics* 6(2). https://doi.org/10.1515/jhsl-2019-1010.

Sinnemäki, Kaius & Di Garbo, Francesca. 2018. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in Psychology* 9. doi:10.3389/fpsyg.2018.01141. https://www.frontiersin. org/article/10.3389/fpsyg.2018.01141.

Spolsky, Bernard & Richard D. Lambert. 2005. Language planning and policy: Models. In Keith Brown (ed.) *Encyclopedia of language and linguistics*. Elsevier.

Stassen, Leon. 1985. *Comparison and Universal Grammar*. Oxford: Basil Blackwell.

Thurston, William R. 1987. Processes of Change in the Languages of North-Western New Britain. *Pacific Linguistics*. B-99. Canberra: Australian National University.

Trudgill, Peter. 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Türker, Emel. 2000. *Turkish-Norwegian codeswitching: Evidence from intermediate and second generation Turkish immigrants in Norway*. Oslo: University of Oslo. (PhD Dissertation.)

Uther, Maria, Knoll, Monja A., & Burnham, Denis. 2007. Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech Communication* 49(1). 2-7.

Vehkalahti, Kimmo. 2019. Kyselytutkimuksen menetelmät ja mittarit [Methods and measures in questionnaire surveys]. Helsinki: University of Helsinki. https://doi.org/10.31885/9789515149817.

Verkerk, Annemarie & Di Garbo, Francesca. Accepted, 2021. *Socio-geographic correlates of typological variation in northwestern Bantu gender systems*. To appear in *Language Dynamics and Change*.

Weatherholtz, Kodi, Campbell-Kibler, Kahryn & Jaeger, T. Florian 2014. Socially-mediated syntactic alignment. *Language Variation and Change* 26(3). 387-420.

Winford, Donald. 2010. Contact and borrowing. In Hickey, Raymond (ed.) *The Handbook of Language Contact*. 170-187. Somerset: John Wiley & Sons, Incorporated.

Wray, Alison & Grace, George W. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117(3). 543-578. https://doi.org/10.1016/j.lingua.2005.05.005.